Multimed Tools Appl (2017) 76:4505–4521
DOI 10.1007/s11042-016-3945-6

Landmark-based multimodal human action recognition

Stylianos Asteriadis[1] · **Petros Daras**[2]

Received: 8 October 2015 / Revised: 24 August 2016 / Accepted: 6 September 2016 /
Published online: 19 September 2016

Abstract Human activity recognition has received a lot of attention recently, mainly thanks to the advancements in sensing technologies and systems' increasing computational power. However, complexity in human movements, sensing devices' noise and person-specific characteristics impose challenges that still remain to be overcome. In the proposed work, a novel, multi-modal human action recognition method is presented for handling the aforementioned issues. Each action is represented by a basis vector and spectral analysis is performed on an affinity matrix of new action feature vectors. Using modality-dependent kernel regressors for computing the affinity matrix, complexity is reduced and robust low-dimensional representations are achieved. The proposed scheme supports online adaptivity of modalities, in a dynamic fashion, according to their automatically inferred reliability. Evaluation on three publicly available datasets demonstrates the potential of the approach.

Keywords Spectral clustering · Human action recognition · Multimodal fusion

✉ Stylianos Asteriadis
stelios.asteriadis@maastrichtuniversity.nl

Petros Daras
daras@iti.gr

[1] Department of Data Science and Knowledge Engineering, University of Maastricht, Postbus 616, 6200 MD, Maastricht, The Netherlands

[2] Information Technologies Institute, Centre for Research & Technology - Hellas, 1st km Thermi - Panorama, 57001, Thessaloniki, Greece

1 Introduction

Human-machine interaction is entering a new era, with computers altering the way they respond to human stimuli. Natural interaction, expressivity, affect [4] and activity recognition [1] are the principal factors that enrich a human-machine interaction experience. Indeed, technology now offers an increasingly large amount of sensing devices for capturing human activity and, in many cases, hidden intentions, behaviors, affective and cognitive states. Wearable inertial measurement sensors [11], robust video processing algorithms [1], infrared and depth sensors [7] and audio [27] are only a few of the cues available for understanding human activity. These advances brought automatic action recognition to the front-end in many applications, ranging from entertainment to health-care systems. Based on the above, it is understood that a robust action recognition scheme should fulfil a series of criteria. First of all, algorithms guaranteeing real time performance are necessary, while accuracy is equally important, especially when it comes to critical circumstances, such as those involving healthcare systems. Although the more information is provided to a system, the more accurate feedback it is likely to deliver, in many circumstances, a large volume of information dramatically increases computational complexity, leading to systems not appropriate for real-time applications. Exploiting multi-modal information is also a significant task that can boost the performance of a system but care should be taken for placing more importance on 'good' modalities than on noisy ones.

In the proposed work, a real-time, human action recognition method is introduced. The proposed framework approaches the problem by taking into account the aforementioned challenges. In particular, a low-dimensional representation of large dimensionality feature vectors is utilized, by following a landmark-based spectral analysis scheme. In this way, low-dimensional subspaces, encoding valuable information, are built, while new, unknown actions are projected on them. Consequently, only valuable information from different modalities is identified and used in the construction of the models and in further classification of new instances. Based on the mathematical framework of spectral analysis, a method for constructing the adjacency matrix combining cues from multiple modalities, is also introduced in this work. Modalities are fused adaptively, according to automatically inferred reliability metrics, guaranteeing increased robustness to sensor's instability or tracking failures. Furthermore, a methodology for catering for large variance within the same action is proposed; in this manner, different styles in executing the same action are handled, boosting, in this way, the system's ability to generalize for unknown individuals. Finally, for inferring for new, unseen vectors, no local sub-manifold unfolding is necessary and, thus, only simple matrix operations are needed, making, thus, the proposed technique suitable for high demands in real time applications. The above are illustrated through experiments, where comparisons with state-of-the-art methods on three datasets are presented (HMMs & Bayes classification, Bag-of-Words used in Support Vector Machines, multiclass Multiple Kernel Learning) and classification speed is assessed.

The proposed technique builds on authors' preliminary work on Microsoft kinect-based activity recognition based on spectral analysis, [3] where results were presented on the single-modality case of only depth data, while inter- and intra-individual sub-actions were not considered and experiments were limited to a single scenario. The rest of the paper is structured as follows: Section 2 gives an overview of systems employed for human action recognition. Section 3 provides the technical details of the proposed method, while Section 4 presents extensive experiments on three publicly available datasets. Section 5 concludes the paper.

2 Related work

Feature pre-processing is strongly related to the utilized cue, in problems related to human activity recognition. Raw inertial sensor data are used extensively, due to their ability to capture instantaneous features of local character and, thus, lead to a rich source of information for action classification. Statistical [23], expressivity [5] and frequency domain parameters [17], on the other hand, although local, convey a summary of an action for different parts of the human body and, thus, they can be time independent. Such parameters usually depend on efficient tracking in video sequences, which is a challenging area of research on its own, attracting the attention of numerous researchers. Recent advances in object tracking have given rise to new techniques aiming at handling (self-)occlusions and local anomalies, using uncertainty-based techniques [36]. Space-Time Volumes [15] concatenate consecutive vision-based two-dimensional human silhouettes along time, leading to three-dimensional volumes and have been extensively used in non-periodic activities, with their performance in the case of varying speed and motion still questioned [1]. Local descriptors (e.g. SIFT [24] and Histograms of Oriented Gradients [19]) necessitate optimal alignment between training and testing data and, although they possess strong discriminative power, they fail to take advantage of whole body actions. A recently proposed approach in the domain of computer vision has introduced the notion of mid-level descriminative patches [12] to automatically extract semantically rich spatial or spatiotemporal windows of RGB information, in order to distinguish elements that account for primitive human actions. Various feature extraction techniques have also been proposed in the area of depth maps for human action recognition; typical is the work in [6], where the authors proposed the use of Depth Motion Maps (DMMs) for capturing motion and shape cues concurrently. Subsequently, LBP descriptors are employed for describing rotation invariant textures of the patches employed. Recently, Song et al. [26] conducted experiments in re-projecting multiple modalities to a new space where correlation among them is maximised and showed that, following this pre-processing step, nonlinear relationships among different data sources can be found.

On a second level lay the methodologies which use as input processed features. The robustness of the selected approach depends on the context of the application and the availability in features. Dynamic Time Warping (DTW) [30] is one of the most well-known classification schemes. One of the major advantages of the method is its adjustability to varying time lengths, but it usually requires a very large number of training examples, as it is basically a template matching technique. Models describing statistical dependencies have also been used extensively, mainly in order to encode time-related correlations. One of the classical approaches, in this vein, are the Hidden Markov Models (HMMs) [16, 35]. Authors in [32], propose a discriminative parameter learning method for a hybrid dynamic network in human activity recognition. They showcase results on walking, jogging, running, hand waving and hand clapping activities. Authors in [20] employ DBNs for the semantic analysis of sports-related events in videos. The probabilistic behavior of human motion-related features has also been widely used through Support Vector Machines (SVMs). SVMs seek hyperplanes in the feature space for separating data into classes. The data points on the margin of the hyperplane are called support vectors. Laptev et al. [18] use non-linear SVMs for the task of recognizing daily activities of small temporal length (answer the phone, sit down/up, kiss, hug, get out of car). Similar, authors in [29] use SVMs on temporal and time-weighted variances, and authors in [21] employ SVMs in RGB and Depth data to recover gestures, and then apply a fusion scheme using inferred motion and audio, in a multimodal environment. Authors in [14] have also utilized SVMs for activity feature classification, on

joint orientation angles and their forward differences, while view-invariant features (normalized between-joint distances orientations and velocities) have been employed in [28]. The output of an Artificial Neural Network (ANN) can also be used for modelling the probability $P(y|x)$ of an activity y to occur, given input feature vector x. Three and four layer perceptrons are among the most common architectures. Typical is the work in [9], where the authors perform indoors action recognition, using two modalities, namely, wearable and depth sensors. Authors in [10] have also recently proposed a method for human action recognition based on skeletal information, using Hierarchical Recurrent Neural Networks, in order to epxloit temporal information in different parts of the human body, while the work in [13] is proposing a three-dimensional Convolution Neural Network in order to jointly make use of spatial and temporal information. Using Neural Networks, special attention should be paid to high complexity during training, as well as overfitting. Classical classification schemes, such as k-Nearest Neighbor-based ones (k-NNs) and binary trees have also been widely reported in the bibliography. The authors in [17] employ Discrete Fourier Transform (DFT) as their representation scheme and feed the corresponding parameters to a k-NN. The main drawbacks of these systems is that they are quite sensitive to parameter fine tuning and tend to generalize poorly for uknown subjects. Recently, there is also a surge in the use of Sparse Representation techniques, especially in the area of computer vision tasks [25, 33, 34], and authors in [37] propose a novel methodology for pattern recognition, applied on action, face, digit and object recognition by transferring the data structure into the optimization process.

3 Landmark-based action recognition

Identical or similar actions represented by feature vectors $\mathbf{x}_i \in \mathbb{R}^m$ can be considered to lay close to each other on a manifold space. Thus, they can be approximated by the linear combination of representation vectors $\mathbf{z}_i \in \mathbb{R}^k$ ($k << m$) with a set of basis vectors $\mathbf{l}_j \in \mathbb{R}^m$, leading to the optimization problem of minimizing $||X - LZ||$, with $X = [\mathbf{x}_1, ..., \mathbf{x}_n] \in \mathbb{R}^{m \times n}$ being a set of n actions, $L = [\mathbf{l}_1, ..., \mathbf{l}_k] \in \mathbb{R}^{m \times k}$ a table of feature vectors corresponding to landmark-features (derived randomly, after clustering or straight from the activities themselves) and $Z = [\mathbf{z}_1, ..., \mathbf{z}_n] \in \mathbb{R}^{k \times n}$ the low-dimensional representation of X. A typical approach for finding low-dimensional representations in manifold spaces is the calculation of distances among all n data vectors, leading to the adjacency matrix $W = (w_{i,j})_{i,j=1}^{n}$ [31]. From W, the degree diagonal matrix D is built, whose elements are the column (or row) sums of W. Subtracting W from D gives the graph Laplacian matrix L, and the eigenvectors corresponding to its k smallest eigenvalues are the low (k)-dimensional representation of the initial dataset. However, large datasets lead to time consuming construction and eigendecomposition of the Laplacian. Moreover, real-time action classification, using a spectral analysis scheme, requires a per-frame unfolding of local submanifolds, as well as the use of a pre-defined number of closest feature points in it. Authors in [8] present a methodology for solving the problem by only using a subset of feature (basis) vectors $\mathbf{l}_j$ instead of finding one-to-one relationships among all feature vectors in a dataset, for building the adjacency matrix. According to this method, the n data points $\mathbf{x}_i \in \mathbb{R}^m$ can be represented by linear combinations of k ($k \ll n$) representative landmarks (basis vectors). This representation can be used in the spectral embedding. The new representations are k-dimensional vectors $\mathbf{b}_i \in \mathbb{R}^k$ while the landmarks are the result of random selection or a k-means algorithm. We hereby extend this technique by introducing a dynamic weighting scheme for handling multiple modalities in the adjacency matrix and provide a framework for real time inference,

using simple matrix operations avoiding, thus, manifold unfolding in testing, which would be prohibitive for real time applications.

Instead of finding representative feature vectors, as in [8], though clustering, here, it is straightforward to extract landmark basis vectors representing whole actions. Each of these k' classes of a training dataset can constitute a basis for building the landmark matrix $L \in \mathbb{R}^{m \times k'}$. Here, we consider each (sub-)action-specific landmark as the average of the corresponding m-dimensional feature vectors. The original data matrix $X = [\mathbf{x}_1, ..., \mathbf{x}_n] \in \mathbb{R}^{m \times n}$ can be approximated by the product of L and the representation matrix $Z \in \mathbb{R}^{k' \times n}$:

$$X \approx LZ \tag{1}$$

Since different individuals (or the same individual, at different times) might adopt different expressivity for performing the same action, the idea of sub-action basis vectors in the spectral embedding is proposed here. In particular, since an action may be defined by more than one classes, a within-action clustering scheme is followed. For a given action a, a hierarchical cluster tree is used, in order to lead to the identification of significant subclusters. The algorithm computes the matrix $Y \in \mathbb{R}^{n_a \times m}$ of the cosine distance between pairs of the n_a feature vectors belonging to the same action. It constructs k_a clusters using the distance criterion, finding the lowest height where a cut through the hierarchical tree leaves a maximum of a pre-defined number of sub-clusters. A stopping criterion is also imposed, so that heavily imbalanced clusters are not created. Using the above, the total number of the landmarks used for spectral classification is $k = \sum_{a=1}^{k'} k_a \geq k'$.

Each element z_{ji} of the representation matrix Z can be found as the output of a kernel function $k_h(\cdot)$ (here, we use the Laplacian Kernel) of feature vector $\mathbf{x}_i$ and landmark $\mathbf{l}_j$ normalized with the sum of the corresponding values for all landmark vectors:

$$z_{ji} = \frac{e^{\frac{-\|\mathbf{x}_i - \mathbf{l}_j\|}{\sigma}}}{\sum_j e^{\frac{-\|\mathbf{x}_i - \mathbf{l}_j\|}{\sigma}}} \tag{2}$$

with $\|\cdot\|$ being a vector distance metric, while σ is the width of the kernel. Z represents the similarity values between data vectors and actions' (or sub-actions') representative landmarks and defines an undirected graph $G = (V, E)$ with graph matrix $W = \hat{Z}^T \hat{Z}$, where:

$$\hat{Z} = D^{-1/2} Z \tag{3}$$

with D being a diagonal matrix whose elements are the row sums of Z. Since each column of the representation matrix sums up to 1, it is straightforward to check that the degree matrix of W is the identity matrix. Consequently [22], the eigenvectors of W are the same as those of the corresponding Laplacian matrix.

Then, the eigenvectors $A = [\mathbf{a}_1 ... \mathbf{a}_k] \in \mathbb{R}^{k \times k}$ and eigenvalues σ_j^2 of $\hat{Z}\hat{Z}^T$ are calculated. It is obvious that σ_j are the singular values of $\hat{Z}$ and A consists of the left singular vectors of $\hat{Z}$, found through singular value decomposition (4), while $B = [\mathbf{b}_1 ... \mathbf{b}_k] \in \mathbb{R}^{n \times k}$ are the eigenvectors of matrix $W = \hat{Z}^T \hat{Z}$. Each row of B is a low-dimensional representation of the original, high-dimensional feature vectors.

$$\hat{Z} = A \Sigma B^T \tag{4}$$

Consequently, and since $A^T = A^{-1}$, B can be computed directly from (4), as:

$$B = (\Sigma^{-1} A^T \hat{Z})^T \tag{5}$$

Σ is a diagonal with elements σ_j, in decreasing order.

3.1 Dynamic fusion of different modalities

The system described above provides an analytical framework that can be easily extended for dynamically fusing different information sources, according to automatically inferred reliability metrics, injected directly into the similarity values between new features and basis vectors. Different modalities may not be equally suitable for the classification problem. Issues attributed to noisy measurements, uncertainties caused by occlusions, or even lack of correlation between a considered input channel and the activities to be detected are factors that, if taken into account during modelling and evaluation, are expected to optimize an action classification scheme performance. In this work, we introduce modality-specific kernel widths σ^c for calculating the representation matrix. When properly weighted, they can adjust the amount of reliability attributed to each modality. This can be achieved by considering that σ^c increases with the probability of model $\theta^{c,f}$ of modality c and feature f generating observation $x^{c,f}$ and is calculated as the normalized average for each modality, using the following equations:

$$p^{c,f} \equiv Pr(X = x^{c,f} | \theta^{c,f}) \tag{6}$$

$$p^c = \frac{1}{N_c} \sum_f p^{c,f} \tag{7}$$

$$\sigma^c = \eta^c \times \frac{p^c}{\sum p^c} \tag{8}$$

N_c is the number of features used for modality c and η^c is a multiplying factor. Thus, (2), for given feature and basis vectors $\mathbf{x}_i^c$, $\mathbf{l}_j^c$, corresponding to modality c, becomes:

$$z_{ji} = \frac{\sum_c e^{\frac{-\|\mathbf{x}_i^c - \mathbf{l}_j^c\|}{\sigma^c}}}{\sum_j \sum_c e^{\frac{-\|\mathbf{x}_i^c - \mathbf{l}_j^c\|}{\sigma^c}}} \tag{9}$$

3.2 Classification of new instances

For classifying a new data vector $\mathbf{x}' = [\mathbf{x}'^1 ... \mathbf{x}'^M]$, coming from M modalities, to an activity, the elements z'_j of the representation vector $\mathbf{z}' \in \mathbb{R}^k$ defined by the similarities between $\mathbf{x}'$ and $L = [(\mathbf{l}_1^1 ... \mathbf{l}_1^M)^T ... (\mathbf{l}_k^1 ... \mathbf{l}_k^M)^T]$ are found as:

$$z'_j = \frac{\sum_c e^{\frac{-\|\mathbf{x}'^c - \mathbf{l}_j^c\|}{\sigma^c}}}{\sum_j \sum_c e^{\frac{-\|\mathbf{x}'^c - \mathbf{l}_j^c\|}{\sigma^c}}} \tag{10}$$

The representation $\mathbf{b}'$ of the new feature vector in the low dimensional domain is given by:

$$\mathbf{b}' = \Sigma^{-1} A^T D^{-1/2} \mathbf{z}' \tag{11}$$

Classification result is given as the label C of the action with low-dimensional representation matrix B_a (as calculated in training) that minimizes a distance metric $d(\cdot)$ from $\mathbf{b}'$:

$$C = \arg\min_a d(\mathbf{b}', B_a) \tag{12}$$

Thus, for new data vectors, no local sub-manifold unfolding is necessary and, for inference, simple matrix operations are needed. This is of great significance, since it allows for real-time action recognition and constitutes the proposed method appropriate for online evaluation of whether the projection of multiple modality features over the course of an action is close to the subspace classes of a trained model.

The overall system is summarized in Algorithm 1:

Algorithm 1 Human action recognition

Input for training:
n data points $\mathbf{x}_1, ..., \mathbf{x}_n$
Their labels
Training:
1: Use hierarchical cluster trees within each action class, separately, to infer landmarks corresponding to possible action subclusters
2: Calculate average feature vectors for each action or sub-action, $\mathbf{l}_1, ..., \mathbf{l}_k$
3: Calculate a reliability kernel for each modality, according to (6)–(8)
4: Calculate representation matrix Z according to (9) and $\hat{Z}$, according to (3).
5: Compute the eigenvectors $A = [\mathbf{a}_1, ..., \mathbf{a}_k]$ and square eigenvalues of ZZ^T
6: Compute B according to (5)
Classify new instances:
1: Online calculation of reliability for each modality
2: Calculate representation vector according to (10)
3: Find low-dimensional representation through (11)
4: Use low-dimensional representation matrix B and labels, in order to infer classification result through (12)

4 Experimental evaluation

In order to have its accuracy validated, the proposed methodology has been tested on three publicly available datasets.

4.1 Skoda Mini Checkpoint Dataset

In the Skoda Mini Checkpoint Dataset, one person, during a 3 hour recording, performed 70 repetitions of 10 activities in a car maintenance scenario (Fig. 1). Motion was captured using 20 accelerometers, placed on the left and right upper and lower arms. Each accelerometer consists of its values on the x, y and z axis. In the experiments, in order to capture temporal and not only qualitative characteristics, every instance was split into 4 periods and the average values of the above features were calculated within these time segments. The above procedure gave a total of 240 features per instance.

For evaluation, the dataset was separated into seven parts of 100 instances, with activities uniformly distributed. For extracting the training matrices, 5 parts were used, while one part

Fig. 1 Example from the Skoda Mini Checkpoint dataset

(validation session) was used for determining the kernel width (2) that can give the highest accuracy. In case of similar accuracies for different kernel widths, the one corresponding to the lowest Sum of Squared Error (SSE) criterion of the low-dimensional classes was used. Following the above procedure, an overall accuracy equal to 98.8 % was achieved. Authors in [35] perform classification using Hidden Markov Models (HMM) on individual nodes. The resulting classifiers are fused by employing a Naive Bayes Classifier, achieving a total of 98 %. Figure 2 summarizes results obtained after extracting landmarks as average feature vectors per activity as well as through random selection and kmeans, similar to [8]. It can be seen that, using as landmarks average feature vectors for each activity, separately (10, in this case), achieves better results than randomly or based on a kmeans algorithm, extracting the same number of landmarks. The last two options gave results comparable to ours, only for a large number of landmarks. However, this comes to a much higher computational cost. Indicatively, a 240-sized feature vector necessitates 0.017 s for classification, when the number of landmarks is equal to 10, while this time becomes four times higher for the double number of landmarks.

4.2 Huawei/3DLife Dataset 1

Experiments on data using non obtrusive equipment were carried out, so as to test the efficacy of the proposed scheme in more noisy but less obstrusive environments. Specifically, using the same set of features as the ones employed in [28], the method was also tested on

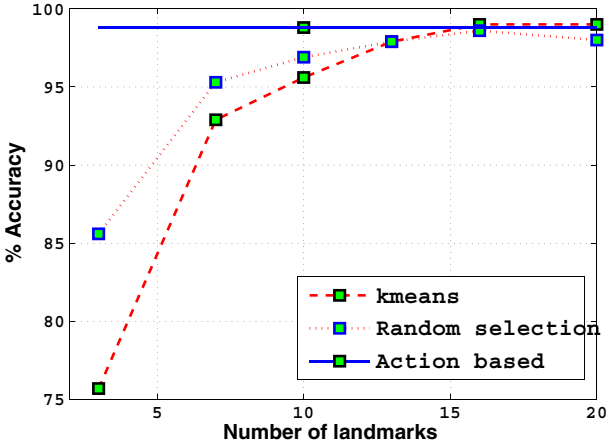

Fig. 2 Activity Recognition rates following landmark selection based on Average Feature Vectors per activity, random selection and kmeans

the Huawei/3DLife Dataset 1, Session 2,[1] where 14 subjects participated, each performing a set of 16 repetitive actions. These actions are either sports-related, or involve some standard movements (e.g. knocking on the door), as shown in Fig. 3. Each action was performed 5 times by each subject. Subjects' motion was captured using a series of depth sensors (Microsoft Kinect). As authors in [28] report results on the non-repetitive action of running on a treadmill, we hereby included this action in our experiments, as well.

Using Kinect depth sensors, human motion can be easily extracted in the form of moving human skeletons [2] and real-time feedback regarding a series of features' positions is obtained (head, neck, shoulders, elbows, hands, torso, hips, knees, feet). Authors in [28] introduce a set of view-invariant features that we hereby present in brevity: For each joint, its distance on all three axis from the torso, (as the torso is seen in the first frame of each action) is calculated. This is normalized with the average distance between the torso joint and the feet joints, in order to cater for different body sizes. Moreover, joint orientations expressed in quaternions are used. Also, velocity information is used, both using positional and orientation-related information. Velocities are calculated for two different time intervals for each feature. The above strategy leads to 264-dimensional features per time segment. Sun and Aizawa [28] use the above features and, after a feature refinement step, they represent them by Bags of Words at sampling intervals of the whole sequence of the action, as well as three temporal subsequences and they use SVM for classification. Similarly, in our experiments, we used the expected values of the same features over the course of each action, as well as, three subsequences of them, which assists in differentiating between similar actions with temporal differences (e.g. backward vs forward tennis moves). Since many actions consist of less then 5 frames, velocity-related features were extracted for two time segments.

Since reliability of each of the above features may vary depending on their origin, 4 different modalities were considered: Raw values of Positions, Raw values of Orientations, Positional Velocity and Orientational Velocity. Using the training data, a distribution

[1] Huawei/3DLife ACM Multimedia Grand Challenge for 2013

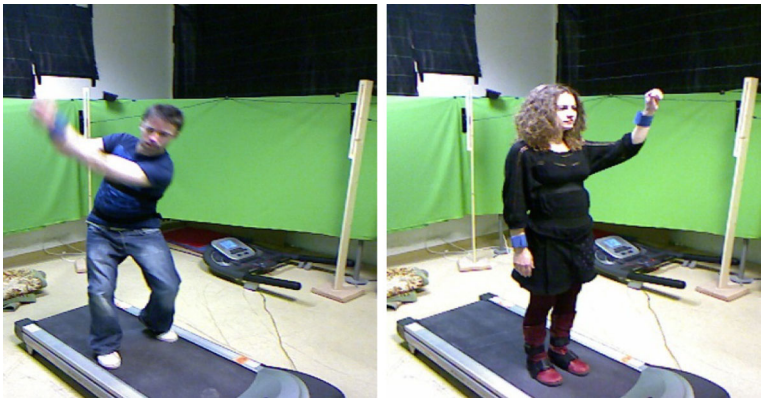

Fig. 3 Examples from the Huawei/3DLife Dataset 1

separately for each feature variable is found and the corresponding new feature variables are expected to fit well in it. In this dataset, gaussian distributions were found to fit well with the data and, as such, reliability for each modality c of feature i can be given by (13):

$$\sigma^c = \eta^c \times \frac{\frac{1}{N_c}\sum_i \frac{1}{\sigma_i^c\sqrt{2\pi}} e^{-\left(\frac{(x_i^c-\mu_i^c)^2}{2\sigma_i^{c\,2}}\right)}}{\sum_j \frac{1}{N_j}\sum_i \frac{1}{\sigma_i^j\sqrt{2\pi}} e^{-\left(\frac{(x_i^j-\mu_i^j)^2}{2\sigma_i^{j\,2}}\right)}} \tag{13}$$

where μ_i^c and σ_i^c are the mean and standard deviation of feature i of modality c and η^c a modality-specific parameter. N_c is the number of feature variables in modality c.

For training, as before, a leave-one-subject out protocol was followed, the Mahalanobis distance was used in (12), while the maximum allowed number of sub-clusters per action was two, and highly imbalanced sub-clusters were merged into the same cluster. Table 1 shows results achieved with the proposed method and different combinations of modalities. It can be seen that, by fusing all feature modalities using proper reliability indicators, accuracy is maximized, while landmark-based action recognition achieves slightly higher results than the popular method relying on Bag-of-Words employed in [28] on the same features. In both experiments, for classification of new feature vectors, less than 0.02 s were necessary, while training for each subject requires about 17 s using non-optimized Matlab code.

Table 1 Results on the Huawei/3DLife Dataset Session 2 using the proposed technique with/without reliability, different combinations of modalities and the technique described in [28]

All modalities (using reliability indicators)	80.4 %
All modalities (not using reliability indicators)	77.6 %
Position and Orientation raw values	71.0 %
Position and Orientation velocities	72.1 %
Method in [28] (Bag-of-words/SVM)	79.78 %

4.3 Berkeley MHAD database

Experimental results are also presented on the recently published dataset, Berkely MHAD (Multimodal Human Action Database), described in [23]. The dataset comprises 11 actions performed by 12 subjects, with each subject performing a set of 5 repetitions of each action. Three different types of actions resulted in a total of 82 min of recording time: 1) actions in both upper and lower body extremities, 2) actions with high dynamics in upper extremities, 3) actions with high dynamics in lower extremities. The actions performed in the dataset, are: jumping, jumping jacks, bending, punching, waving two hands, waving one hand, clapping, throwing, sit down/stand up, sit down, stand up. For each action, 5 different cues were used for recognition: A Mocap System, a set of multi-view video data, a set of two Microsoft Kinect depth sensors, six three-axis accelerometers that capture the motion of hips, ankles and wrists, and an audio system.

For the experiments in the proposed work, a set of 12 joint angles were used, as calculated from the mocap data (see Fig. 4). Their variance in 5 successive temporal windows was calculated, for each action. The above procedure led to a total of 180 features per action. All accelerometer data were employed (6 three-dimensional vectors), and their variances in 15 temporal windows was considered, leading to a total of 270 features per action. Similar to [23], the 7 first subjects were used for training, while the last 5 were used for testing.

As explained in Section 3.1, for efficient fusion of the two cues, reliability metrics must be established. Here, using the training data, the distribution for each feature variable is found and new features are expected to fit well in it. Subsequently, the probability density function values of this variable for new features is calculated. The distribution considered in this case, for each feature, is the lognormal. More in particular, it was noticed that the data corresponding to each feature variable do not exhibit symmetry but, instead,

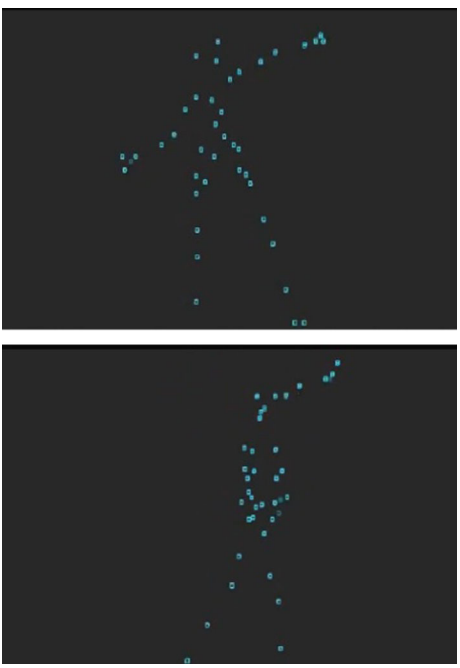

Fig. 4 Examples from the Motion Capture data, during the action of "Throwing" [23]

Table 2 Proposed method and method in [23] results on mono-modal and multi-modal instances of the Berkeley/MHAD dataset

	Proposed method	[23]
MOCAP data	84.0 %	79.9 %
Accelerometer	72.0 %	85.4 %
MOCAP + Accel.	98.18 %	97.45 %

their distributions have large skews towards the positive direction and small skews towards the negative one. Consequently, the common choice of a gaussian distribution should be avoided. Instead, in this case, opting for lognormal parameterizations is more straightforward. Thus, each feature i, belonging to modality c, is considered to follow a lognormal distribution $f(x_i^c \mid \mu_i^c, \sigma_i^c)$ with μ_i^c and σ_i^c being the mean and standard deviation, respectively, of the associated normal distribution. Equation (14) can then be used to obtain the normalized weight corresponding to each modality c:

$$\sigma^c = \eta^c \times \frac{\frac{1}{N_c}\sum_i \frac{1}{\sigma_i^c\sqrt{2\pi}} e^{-\left(\frac{(ln(x_i^c)-\mu_i^c)^2}{2\sigma_i^{c\,2}}\right)}}{\sum_j \frac{1}{N_j}\sum_i \frac{1}{\sigma_i^j\sqrt{2\pi}} e^{-\left(\frac{(ln(x_i^j)-\mu_i^j)^2}{2\sigma_i^{j\,2}}\right)}} \tag{14}$$

with η^c being a modality specific constant and N_c the number of feature variables in modality c. In our experiments, η^c was set to 6, for both modalities, as it achieved the best

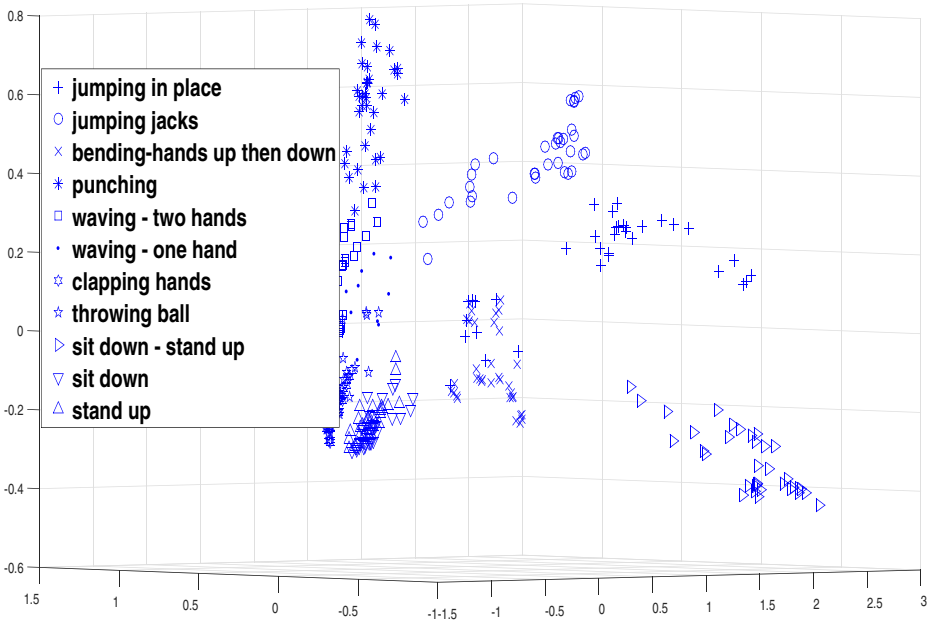

Fig. 5 Action classes represented by the 3 elements of $\mathbf{b}_j$ explaining the highest variance among features, for the Motion Capture features in the Berkeley-MHAD dataset

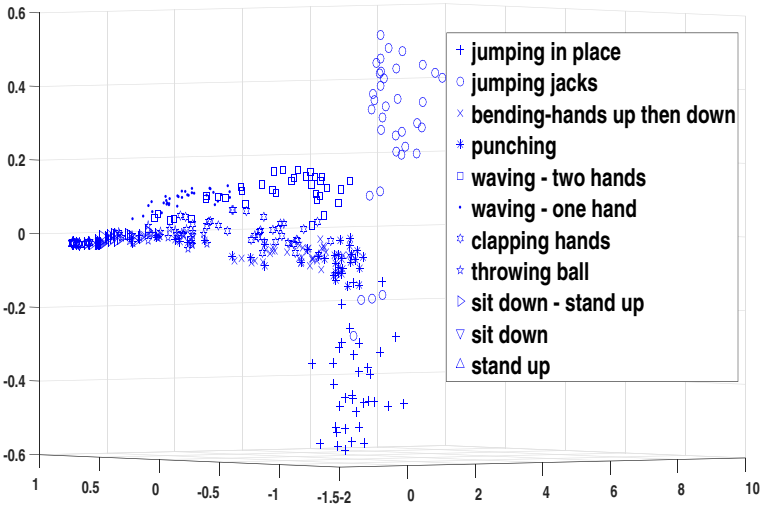

Fig. 6 Action classes represented by the 3 elements of $\mathbf{b}_j$ explaining the highest variance among features, for the Accelerometer features in the Berkeley-MHAD dataset

accuracy on a validation dataset of 2 subjects, part of the training data of the 7 subjects. Table 2 compares the results achieved using the proposed method and the method used in [23], where multiclass Multiple Kernel Learning was used, while, Figs. 5, 6 and 7 are indicative of the discriminative power of the proposed technique. Specifically, as the corresponding results suggest, using both modalities clearly helps to better distinguish classes from each other that, using one modality alone, would not be possible. Moreover, classes similar to each other (sit down - stand up) can be effectively separated at dimensionalities

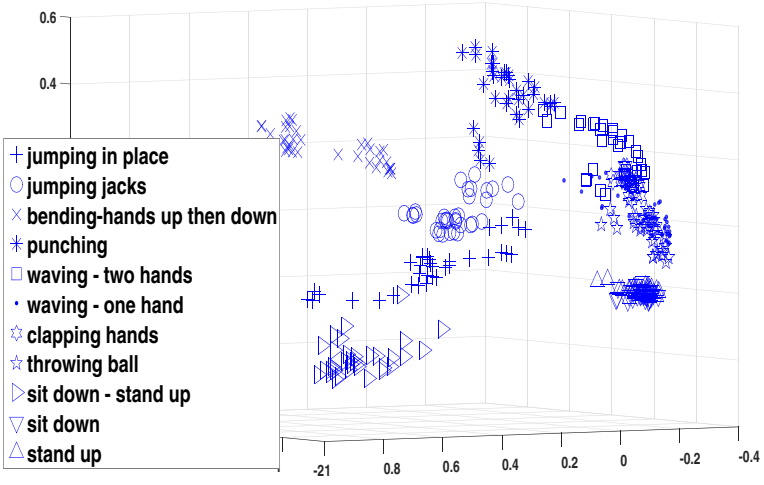

Fig. 7 Action classes represented by the 3 elements of $\mathbf{b}_j$ explaining the highest variance among features, for the fusion of Motion Capture and Accelerometer features in the Berkeley-MHAD dataset

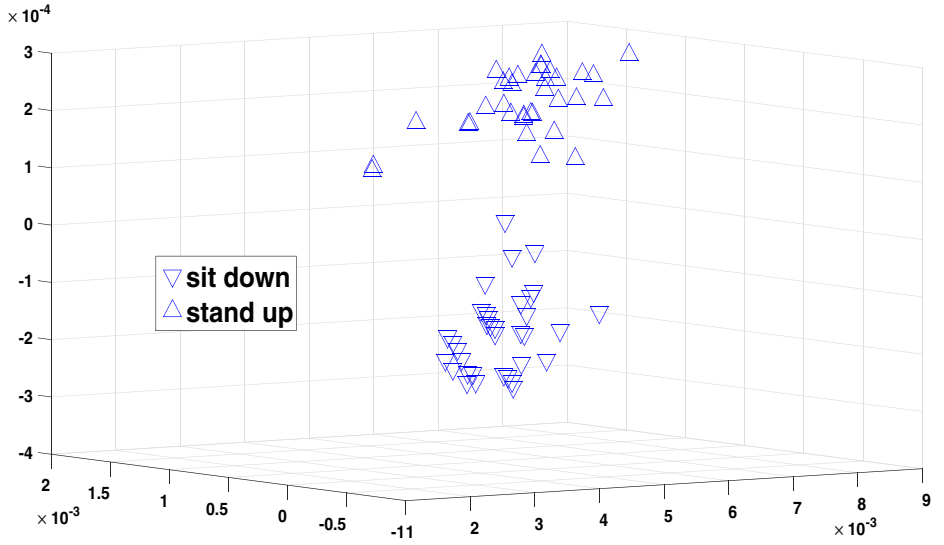

Fig. 8 Stand up - Sit down classes separated by lower-level elements of $\mathbf{b}_j$ for the fusion of Motion Capture and Accelerometer features in the Berkeley-MHAD dataset

of $\mathbf{b}_j$ explaining lower feature variances (Fig. 8). For classification of new feature vectors, less than 0.02 s were necessary, while training on the first 7 subjects requires about 25 s using, non-optimized Matlab code.

5 Conclusions

In this paper, we used action-dependent basis vectors for projecting large-dimensionality feature vectors to low-dimensional spaces. An affinity matrix between feature and basis vector was constructed, instead of the full adjacency matrix. In the proposed method, catering for different action styles is taken into consideration, while, an online, adaptative, weighting modality scheme is proposed in the representation matrix. Evaluation on three publicly available datasets showed that the method is promising and that the proposed technique, building on multimodal spectral analysis, can achieve high levels of accuracy, comparable or even higher than techniques using state of the art methods in the field (Bag of Words, Hidden Markov Models, Support Vector Machines). Moreover, the proposed method provides with an analytical approach for action recognition, using expressivity-dependent features. This can alleviate from constraints imposed by the Markovian assumption in HMMs and the large number of training data that need to be used. Finally, as seen through experiments, the method can be used for real-time applications, since simple matrix operations are needed for inference; for our classification purposes, in each of the experiments, less than 0.02 s were needed for each instance, using non-optimized code, which is a promising result for on-the-fly recognition of activities in a multimodal environment.

Acknowledgments This work has been partly funded by the EU Horizon 2020 Framework Programme under grant agreement no. 690090 (ICT4Life project).

References

1. Aggarwal JK, Ryoo MS (2011) Human activity analysis: a review. ACM Comput Surv 43(3):16
2. Asteriadis S, Chatzitofis A, Zarpalas D, Alexiadis DS, Daras P (2013) Estimating human motion from multiple kinect sensors. In: Proceedings of the 6th international conference on computer vision/computer graphics collaboration techniques and applications, p 3. ACM
3. Asteriadis S, Daras P (2015) Skeleton-based human action recognition using basis vectors. In: International conference on pervasive technologies related to assistive environments (PETRA)
4. Asteriadis S, Karpouzis K, Kollias SD (2008) A neuro-fuzzy approach to user attention recognition. In: 18th international conference on artificial neural networks (ICANN). Prague, 3–6 September 2008, pp 927–936
5. Caridakis G, Castellano G, Kessous L, Raouzaiou A, Malatesta L, Asteriadis S, Karpouzis K (2007) Expressive faces, gestures and speech in multimodal affective analysis. In: Boukis C, Pnevmatikakis A, Polymenakos L (eds) Artificial intelligence and innovations: from theory to applications, pp 375–388
6. Chen C, Liu M, Zhang B, Han J, Jiang J, Liu H 3d action recognition using multi-temporal depth motion maps and fisher vector
7. Chen L, Wei H, Ferryman JM (2013) A survey of human motion analysis using depth imagery. Pattern Recogn Lett 34(15):1995–2006
8. Chen X, Cai D (2011) Large scale spectral clustering with landmark-based representation. In: AAAI conference on artificial intelligence
9. Delachaux B, Rebetez J, Perez-Uribe A, Mejia HFS (2013) Indoor activity recognition by combining one-vs.-all neural network classifiers exploiting wearable and depth sensors. In: Lecture notes in computer science, pp 216–223
10. Du Y, Wang W, Wang L (2015) Hierarchical recurrent neural network for skeleton based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1110–1118
11. He W, Guo Y, Gao C, Li X (2012) Recognition of human activities with wearable sensors. EURASIP J Adv Sig Proc 2012:108
12. Jain A, Gupta A, Rodriguez M, Davis LS (2013) Representing videos using mid-level discriminative patches. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2571–2578
13. Ji S, Xu W, Yang M, Yu K (2013) 3d convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell 35(1):221–231
14. Kapsouras I, Nikolaidis N (2014) Action recognition on motion capture data using a dynemes and forward differences representation. J Vis Commun Image Represent 25(6):1432–1445
15. Ke Y, Sukthankar R, Hebert M (2007) Spatio-temporal shape and flow correlation for action recognition. In: 7th international workshop on visual surveillance
16. Kim E, Helal S, Cook D (2010) Human activity recognition and pattern discovery. IEEE Pervasive Comput 9(1):48–53. doi:10.1109/MPRV.2010.7
17. Kumari S, Mitra SK (2011) Human action recognition using dft. In: Computer vision, pattern recognition national conference on image processing and graphics, vol 0, pp 239–242
18. Laptev I, Marszałek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. In: IEEE conference on computer vision & pattern recognition (CVPR)
19. Lu WL, Little JJ (2006) Simultaneous tracking and action recognition using the pca-hog descriptor. In: The 3rd Canadian conference on computer and robot vision, p 6
20. Luo Y, Wu TD, Hwang JN (2003) Object-based analysis and interpretation of human motion in sports video sequences by dynamic bayesian networks. Comput Vis Image Underst 92(2–3):196–216
21. Nandakumar K, Wan KW, Chan SMA, Ng WZT, Wang JG, Yau WY (2013) A multi-modal gesture recognition system using audio, video, and skeletal joint data. In: Proceedings of the 15th ACM on International conference on multimodal interaction, pp 475–482. ACM
22. Ng AY, Jordan MI, Weiss Y (2001) On spectral clustering: analysis and an algorithm. In: Advances in neural information processing systems. MIT Press, pp 849–856
23. Ofli F, Chaudhry R, Kurillo G, Vidal R, Bajcsy R (2013) Berkeley mhad: a comprehensive multimodal human action database. In: IEEE workshop on applications of computer vision, vol 0, pp 53–60

24. Scovanner P, Ali S, Shah M (2007) A 3-dimensional sift descriptor and its application to action recognition. In: Proceedings of the 15th international conference on multimedia, MULTIMEDIA '07. ACM, New York, pp 357–360
25. Shen C, Chen L, Priebe CE (2015) Sparse representation classification beyond l1 minimization and the subspace assumption. arXiv preprint arXiv:1502.01368
26. Song Y, Morency LP, Davis R (2012) Multimodal human behavior analysis: learning correlation and interaction across modalities. In: Proceedings of the 14th ACM international conference on multimodal interaction. ACM, pp 27–30
27. Stork J, Spinello L, Silva J, Arras K (2012) Audio-based human activity recognition using non-markovian ensemble voting. In: IEEE international workshop on robots and human interactive communications (RO-MAN), pp 509–514
28. Sun L, Aizawa K (2013) Action recognition using invariant features under unexampled viewing conditions. In: Proceedings of the 21st ACM international conference on multimedia, MM '13. ACM, New York, pp 389–392
29. Vantigodi S, Babu RV (2013) Real-time human action recognition from motion capture data. In: 2013 fourth national conference on computer vision, pattern recognition, image processing and graphics (NCVPRIPG). IEEE, pp 1–4
30. Veeraraghavan A, Member S, Roy-chowdhury AK (2005) Matching shape sequences in video with applications in human movement analysis. IEEE Trans Pattern Anal Mach Intell 27:1896–1909
31. von Luxburg U (2007) A tutorial on spectral clustering. Stat Comput
32. Wang X, Ji Q (2012) Learning dynamic bayesian network discriminatively for human activity recognition. In: Proceedings of the 21st international conference on pattern recognition (ICPR), pp 3553–3556
33. Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y (2009) Robust face recognition via sparse representation. IEEE Trans Pattern Anal Mach Intell 31(2):210–227
34. Yang AY, Zhou Z, Balasubramanian AG, Sastry SS, Ma Y (2013) Fast-minimization algorithms for robust face recognition. IEEE Trans Image Process 22(8):3234–3246
35. Zappi P, Lombriser C, Stiefmeier T, Farella E, Roggen D, Benini L, Tröster G (2008) Activity recognition from on-body sensors: accuracy-power trade-off by dynamic sensor selection. Springer
36. Zhang B, Perina A, Li Z, Murino V, Liu J, Ji R (2016) Bounding multiple gaussians uncertainty with application to object tracking. Int J Comput Vis 1–16
37. Zhang B, Perina A, Murino V, Del Bue A (2015) Sparse representation classification with manifold constraints transfer. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4557–4565

Stylianos Asteriadis, is Assistant Professor at the University of Maastricht. Stylianos Asteriadis received his PhD from the School of Electrical and Computer Eng. of National Technical University of Athens (2011), and his Master's and Diploma degrees from Aristotle University of Thessaloniki (2006 and 2004, respectively). He has extensive experience in research on visual and affective computing, machine learning and human-computer interaction, especially in the areas of human activity recognition, emotion analysis, multimodal user interfaces and personalization. Stylianos Asteriadis is a reviewer for several journals and conferences in his field. He is the author of more than 30 journal and international conference papers in the aforementioned fields, while he frequently participates in international research projects (funded by the EU or by national resources) that bring together research and industry.

Petros Daras is a Principal Researcher Grade A', at the Information Technologies Institute (ITI) of the Centre for Research and Technology Hellas (CERTH). He has been working as a Researcher since July 2006 and he has been involved in more than 30 projects, funded by the EC and the Greek Ministry of Research and Technology. His main research interests include multimedia processing, multimedia & multimodal search engines, 3D reconstruction from multiple sensors, dynamic mesh coding, medical image processing and bioinformatics. He has co-authored more than 40 papers in refereed journals, 29 book chapters and more than 100 papers in international conferences. He has served as a regular reviewer for a number of international journals and conferences. He was the chair of the IEEE Interest Group (IG) on Image, Video and Mesh Coding (2012–2014) and key member of the IEEE IG on 3D Rendering, Processing and Communications (2010 -). He regularly acts as a reviewer/evaluator for the EC. He is the head of the Visual Computing Lab. He is a Senior Member of IEEE.